The quality of the algorithm for proven and young with various sets of core animals in a multibreed sheep population¹

Mohammad Ali Nilforooshan^{2,3,✉} and Michael Lee

Department of Mathematics and Statistics, University of Otago, Dunedin 9054, New Zealand

ABSTRACT: The inverses of the pedigree and genomic relationship matrices (**A**, **G**) are required for single-step GBLUP (ssGBLUP). While, inverting **A** is possible for millions of animals at a linear cost, inverting **G** has a cubic cost and feasible for at most 150,000 animals, using the current conventional algorithms. The algorithm for proven and young (APY) provides approximations of the regular ssGBLUP by splitting genotyped animals into core and noncore groups, with computational costs being cubic for core and linear for noncore animals. The data consisted of 9,406,096 animals in the pedigree, 6,243,753 weaning weight phenotypes, and 46,949 genotyped animals from 5 breeds, composites, and animals with missing breed information from New Zealand. Aiming to find a core sample for a multibreed sheep population that can provide evaluations similar to those from the regular ssGBLUP, different core types, and core sizes were studied. Core types random, composite, oldest, youngest, the most inbred animals in **G** (GINB), and in **A** (AINB) were studied in 5K, 10K, and 20K core sizes ($K = 1,000$). Romney core was studied in 5K and 10K, and Coopworth-Perendale core was studied in 5K. Correlation and regression coefficient

(slope) between GEBV from the non-APY and the APY analyses, as indicators for consistency with non-APY and bias from non-APY, showed a large impact of APY on noncore and a small impact on nongenotyped animals. Breed-based 5K cores resulted in large bias from non-APY even for nongenotyped animals. Random and GINB at 20K core size resulted in the highest consistency with non-APY and the lowest bias from non-APY. However, GINB did not perform as well as Random at lower core sizes. The number of animals from a breed in the core sample was very important for the evaluation of that breed. We observed that cores without Texel or Highlander animals resulted in poor evaluations for those breeds. Solving the mixed model equations, within core type, the smallest core size, and within core size, Random core converged in the least number of iterations. However, APY *per se* did not necessarily reduce the solving time. Random cores performed the best, as they could give a good coverage on the generations and breeds, representative for the genotyped population. Core size 20K performed better than 5K and 10K, and the optimum core size was found to be 18.8K, according to the eigenvalue decomposition of **G**.

Key words: APY; BLUP; genomic relationship matrix; inversion; sheep; ssGBLUP

© The Author(s) 2019. Published by Oxford University Press on behalf of the American Society of Animal Science. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

J. Anim. Sci. 2019.97:1090-1101

doi: 10.1093/jas/skz010

¹The authors thank Beef+Lamb NZ Genetics and seed-stock breeders for providing data and funding this research as part of a project for the implementation of the single-step GBLUP evaluation of sheep in New Zealand. The authors also thank Dr. Ismo Strandén and Dr. Martin Lidauer for their technical support for MiX99 software package, Dr. Hossein Jorjani for helpful feedback on the

manuscript, and anonymous reviewers for their constructive comments.

²Present address: Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand

³Corresponding author: m.a.nilforooshan@gmail.com

Received August 30, 2018.

Accepted January 7, 2019.

INTRODUCTION

Single-step genomic BLUP (ssGBLUP) is becoming a popular method for genomic evaluation of livestock. It enables simultaneous evaluation of genotyped and nongenotyped animals. In comparison with BLUP, the inverse of the pedigree relationship matrix ($\mathbf{A}^{-1}$) is replaced with the inverse of an augmented pedigree and genomic relationship matrix (Aguilar et al., 2010; Christensen and Lund, 2010). Henderson (1976) developed an algorithm for indirect inversion of $\mathbf{A}$, which was further improved by Quaas (1976). That method allowed the inverse of $\mathbf{A}$ to be constructed directly at a linear rather than a cubic computational cost, relative to the number of animals, which made BLUP evaluations feasible for large number of animals despite the limited computational power available at the time. Nowadays, the method is used to construct $\mathbf{A}^{-1}$ for millions of animals. The inverses of the genomic relationship matrix ($\mathbf{G}^{-1}$) and the partition of the pedigree relationship matrix for genotyped animals ($\mathbf{A}_{22}^{-1}$) are also required in ssGBLUP, which cannot be inverted indirectly. Inverting $\mathbf{G}$ directly takes quadratic memory and cubic computational time and complexity (Misztal, 2016). Due to computational limitations, inverting $\mathbf{G}$ for more than 150,000 genotypes is currently not practical, using the current conventional algorithms (Fragomeni et al., 2015a, 2015b; Misztal, 2016; Bradford et al., 2017). Misztal et al. (2014) developed the Algorithm for Proven and Young (APY), in which $\mathbf{G}$ is partitioned into proven (core) and young (noncore) animals, and direct inversion is only required for the block of $\mathbf{G}$ corresponding to proven animals. The algorithm reduces computational costs to linear in young animals (Misztal et al., 2014). Later studies showed that the use of proven animals as core and young animals as noncore was unnecessary and that assigning different groups of animals (including random) to core produced genomic EBV (GEBV) similar to GEBV with proven animals as core (Fragomeni et al., 2015a, 2015b; Bradford et al., 2017). The choice of core animals is important for the accuracy of GEBV (Ostersen et al., 2016). Strategies of choosing core animals have been tested in cattle (Lourenco et al., 2015; Fragomeni et al., 2015a, 2015b; Masuda et al., 2016; Bradford et al., 2017), and in 3 breeds of pig (Ostersen et al., 2016). Except a recent study (Vandenplas et al., 2018) on a simulated data with a crossbreeding scheme similar to pigs and chickens, the choice of core animals in APY for a multibreed population and the impact on different groups of animals including, core,

noncore, and nongenotyped animals has not been addressed in the literature. The aim of this study was to evaluate the consistency of APY GEBV predictions with those obtained by the ssGBLUP using the direct inverse of $\mathbf{G}$ (non-APY), in a multibreed New Zealand sheep population. Different core sizes and core types were tested.

MATERIALS AND METHODS

Data

Data were obtained from Beef+Lamb NZ Genetics (Dunedin, New Zealand), in June 2017. It consisted of lamb weaning weights with an average weaning age of 95 d, from a mixture of pure and crossbred animals from several breeds, born between 1997 and 2016. After initial edits and quality controls on the genotype data, a set of 40,132 SNP from Illumina Ovine 50K (Illumina, Inc., San Diego, CA) or LD genotypes imputed (Sargolzaei et al., 2014) to 50K Illumina remained for 46,949 animals, of which 45,477 had phenotypes. In total, there were 9,406,096 animals with pedigree and 6,243,753 phenotypes contributed from 1,050 flocks. Table 1 shows the frequency of breeds in the pedigree, phenotype and genotype data.

Model and Methods

The ssGBLUP method (Aguilar et al., 2010; Christensen and Lund, 2010) was used, in which, the inverse of the covariance matrix ($\mathbf{A}^{-1}$) in BLUP is replaced with $\mathbf{H}^{-1}$, to predict GEBV:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{A}$ and $\mathbf{G}$ are the pedigree and the genomic relationship matrices, and $\mathbf{A}_{22}$ is a submatrix of $\mathbf{A}$ corresponding to genotyped animals. $\mathbf{G}$ was created according to method 1 of VanRaden (2008),

Table 1. Number of animals, phenotypes, and genotypes per breed

Breed	Pedigree	Phenotypes	Genotypes
Composite	1,188,715	879,917	22,800
Coopworth	768,389	598,968	3,093
Highlander	96,968	78,197	503
Perendale	695,585	495,124	2,828
Romney	3,363,136	2,527,208	14,735
Texel	318,833	225,941	760
Unknown	2,974,470	1,438,398	2,230
Total	9,406,096	6,243,753	46,949

and it was scaled by multiplying by $\text{tr}(\mathbf{A}_{22})/\text{tr}(\mathbf{G})$. Furthermore, $\mathbf{G}$ was replaced by $0.95 \mathbf{G} + 0.05 \mathbf{A}_{22}$ to ensure the positive definiteness of $\mathbf{G}$ (Fragomeni et al., 2015b; Misztal et al., 2016).

In one ssGBLUP analysis, the direct-inverse of $\mathbf{G}$ was used (non-APY), and in the other analyses the inverse of $\mathbf{G}$ was obtained using the APY method ($\mathbf{G}_{APY}^{-1}$; Misztal et al., 2014), where direct inversion was only needed for the block of $\mathbf{G}$ corresponding to the core animals. Partitioning $\mathbf{G}$ to blocks corresponding to core (c) and noncore (n) animals:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{G}_{nn} \end{bmatrix},$$

$\mathbf{G}_{APY}^{-1}$ is obtained as (Misztal et al., 2014):

$$\mathbf{G}^{-1} \approx \mathbf{G}_{APY}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1} \mathbf{G}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{G}_{nc} \mathbf{G}_{cc}^{-1} & \mathbf{I} \end{bmatrix}.$$

$\mathbf{M}_{nn}$ is a diagonal matrix of genomic Mendelian sampling with diagonal elements:

$$m_{ii} = g_{ii} - \mathbf{g}_{ic} \mathbf{G}_{cc}^{-1} \mathbf{g}_{ci},$$

where g_{ii} is the i th diagonal element in $\mathbf{G}_{nn}$. Obtaining $\mathbf{G}_{cn}$ and $\mathbf{G}_{cc}^{-1}$ are computationally intensive.

The following model was used to analyze the data, and the mixed model equations (MME) were solved using MiX99 software (Strandén and Lidauer, 1999):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wm} + \mathbf{e},$$

where $\mathbf{y}$, $\mathbf{b}$, $\mathbf{u}$, $\mathbf{m}$, and $\mathbf{e}$, respectively are vectors of phenotypes, fixed effects, direct additive genetic effects, maternal additive genetic effects, and residuals. $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$ are incidence matrices linking $\mathbf{y}$ to $\mathbf{b}$, $\mathbf{u}$, and $\mathbf{m}$. The random effects were distributed as $\mathbf{u} \sim N(0, \mathbf{H}\sigma_u^2)$, $\mathbf{m} \sim N(0, \mathbf{H}\sigma_m^2)$, and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, with $\sigma_u^2 = 2.45$, $\sigma_m^2 = 1.5925$, and $\sigma_e^2 = 9.8$.

Choice of Core Animals

The sampling of core animals was investigated by selecting animals at random (RNDM); based on breed [Romney (ROM), Coopworth-Perendale (CP), Composites (COMP)]; age [oldest (OLD), youngest (YNG)]; and inbreeding [highest in $\mathbf{A}_{22}$ (AINB), and highest in $\mathbf{G}$ (GINB)]. The core size was 5K, 10K, or 20K (K = 1,000). From here on, core size and core type combinations will be

presented as size_type (e.g., 5K_ROM). We considered GINB to study if using a core with the highest diagonal elements of $\mathbf{G}$ can provide a better APY-inverse of $\mathbf{G}$, and as a result, GEBV closer to those from the non-APY analysis. Because creating $\mathbf{G}$ before assigning core animals is expensive, AINB was considered to study if selecting core animals based on the highest diagonal elements of $\mathbf{A}_{22}$ can perform like GINB. Diagonal elements of $\mathbf{G}$ and $\mathbf{A}_{22}$ were in a range of 0.9071 and 1.4717, and 1 and 1.3412, respectively. Selecting 5K, 10K, and 20K animals with the highest diagonal elements in $\mathbf{G}$, the minimum diagonal elements of $\mathbf{G}$ were 1.061, 1.0318, and 1.0096, respectively. Selecting 5K, 10K, and 20K animals with the highest diagonal elements in $\mathbf{A}_{22}$, the minimum diagonal elements of $\mathbf{A}_{22}$ were 1.0385, 1.0200, and 1.0027, respectively. Diagonal elements of $\mathbf{G}$ and $\mathbf{A}_{22}$ had a correlation of 0.404. Table 2 describes the cores, with respect to size and type. The APY analyses involving random core sampling (Table 2), were replicated several times proportionally to the number of core candidates divided by the core size. This was done to account for the variation introduced by random sampling.

Table 2. Description of the cores in the APY analyses

Core	Description
5K_CP	2,500 random Coopworth and 2,500 random Perendale
5K_ROM	5,000 random Romney
5K_COMP	5,000 random Composites
5K_OLD	5,000 oldest ^a
5K_YNG	5,000 youngest ^a
5K_AINB	5,000 of the most inbred in $\mathbf{A}_{22}$ ^b
5K_GINB	5,000 of the most inbred in $\mathbf{G}$ ^c
5K_RNDM	5,000 random
10K_ROM	10,000 random Romney
10K_COMP	10,000 random Composites
10K_OLD	10,000 oldest
10K_YNG	10,000 youngest
10K_AINB	10,000 of the most inbred in $\mathbf{A}_{22}$
10K_GINB	10,000 of the most inbred in $\mathbf{G}$
10K_RNDM	10,000 random
20K_COMP	20,000 random Composites
20K_OLD	20,000 oldest
20K_YNG	20,000 youngest
20K_AINB	20,000 of the most inbred in $\mathbf{A}_{22}$
20K_GINB	20,000 of the most inbred in $\mathbf{G}$
20K_RNDM	20,000 random

^aBecause of the unavailability of exact birth dates, selecting animals based on age was involved selecting animals from a few years and random sampling of animals born in a specific year.

^b $\mathbf{A}_{22}$ = the block of the pedigree relationship matrix corresponding to genotyped animals.

^c $\mathbf{G}$ = genomic relationship matrix.

Comparisons

Comparisons were made for 3 groups of animals: nongenotyped (**nongeno**), genotyped animals used in the core (**core**), and genotyped animals not used in the core (**noncore**). Correlations and regression coefficients of the APY GEBV to the non-APY GEBV were evaluated as indicators of consistency with non-APY and bias from non-APY. The bias from non-APY was defined as the deviation of the slope of the regression line from 1. To study the lack or the shortage of animals from a specific breed in the core sample on the evaluations of animals from that breed, Texel and Highlander breeds were chosen. There were 755 and 503 genotyped animals from Texel and Highlander breeds, respectively (Table 1). We also studied the eigenvalue decomposition of $\mathbf{G}$ (using RSpectra package: Qiu and Mei, 2018), diagonal elements of $\mathbf{G}^{-1}$, sparsity of $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$, distribution of GEBV, number of preconditioned conjugate gradient (PCG) iterations, and time spent to solve MME (convergence criterion = $1.0\text{e-}12$; criterion based on the relative difference between right-hand and left-hand side of the MME, considering all equations). Data handling and the comparisons were made using R statistical software (R Core Team, 2015).

RESULTS

The population structure of genotyped animals is illustrated by plotting the 2 eigenvectors corresponding to the highest eigenvalues of $\mathbf{G}$ against each other. Figure 1 shows the population structure for

Figure 1. Scatter plot of the 2 eigenvectors corresponding to the highest eigenvalues of the genomic relationship matrix (animals with known breed are presented).

purebred animals, and Supplementary Fig. 1 demonstrates Composite and unknown-breed animals. In agreement with previous findings (Fragomeni et al., 2015a; Bradford et al., 2017), the solutions of different replicates were almost identical. Therefore, only results from the first replicate are presented.

Consistency with Non-APY

Figure 2 shows the GEBV correlations between the non-APY and the APY analyses for core, non-core, and nongeno animals. The GEBV correlations were the highest for nongeno and the lowest for noncore animals, greater with a larger core size. There were cases that the core type was more influential on the correlation than the core size. For example, 10K_RNDM performed better than 20K_COMP (Fig. 2). The cores RNDM and GINB were the most correlated with non-APY, and breed-based cores (including COMP) were the least correlated with non-APY. The core OLD gave a higher correlation than YNG, especially for noncore animals (Fig. 2). The gain in the correlation was not linear by core size. It was higher from 5K to 10K than from 10K to 20K (Fig. 2). Some cores not only resulted in inconsistent APY for noncore animals, but also inconsistent APY for core animals. For example, ROM, CP, and COMP cores resulted in poor correlation for core animals (Fig. 2).

Bias from Non-APY

Figure 3 shows the slope of the APY GEBV regressed on the non-APY GEBV. Deviation of coefficients from 1 represents bias from non-APY. Because of the large sample sizes, even small deviations from 1 were significant ($P < 0.005$). For core animals, except for 5K_YNG, all the slopes were greater than 1, showing downward bias. On the other hand, for noncore and nongeno animals, most of the slopes were less than 1, showing upward bias. For nongeno and core animals, 20K_RNDM and 20K_GINB, and for noncore animals, 10K_YNG and 20K_GINB were the least biased (Fig. 3).

The bias from non-APY of 5K_CP was large for the 3 groups of animals (Fig. 3). This was due to the small proportion of genotypes from Coopworth (6.6%) and Perendale (6.0%), and the genotypes from these breeds could not provide valuable core information for the whole population. For genotyped animals (core and noncore), OLD was more biased than YNG (Fig. 3). Part of the similarities between cores was due to the overlap that those had with each other (Supplementary Fig. 2).

Figure 2. Correlations between the non-APY GEBV and the APY GEBV using different cores^a. ^assGBLUP = single-step GBLUP, CP = Coopworth-Perendale, ROM = Romney, COMP = Composite, OLD = Oldest, YNG = Youngest, AINB = most inbred in A_{22} (the block of the pedigree relationship matrix corresponding to genotyped animals), GINB = most inbred in G (genomic relationship matrix), RNDM = Random.

Figure 3. Regression coefficients (slopes) of the APY GEBV using different cores on the non-APY GEBV^a. ^assGBLUP = single-step GBLUP, CP = Coopworth-Perendale, ROM = Romney, COMP = Composite, OLD = Oldest, YNG = Youngest, AINB = most inbred in A_{22} (the block of the pedigree relationship matrix corresponding to genotyped animals), GINB = most inbred in G (genomic relationship matrix), RNDM = Random.

The intercepts were small negative, ranged from -0.140 to -0.002 , -1.334 to -0.009 , and -0.782 to -0.021 for nongeno, core, and noncore animals, respectively (results not shown). The closest intercepts to 0 were for 20K_RNDM and 20K_GINB, consistent with less bias using these cores (Figs 2 and 3).

Impact on Breeds Underrepresented in the Core Sample

The consistency with non-APY and bias from non-APY were studied in Texel and Highlander breeds. Both parameters were influenced by the size and type of the core chosen. For nongeno Texel and Highlander animals, the correlation was close to 1. For genotyped animals, the correlation had a larger decrease for Highlander breed than for Texel breed (Fig. 4). For Texel and Highlander breeds, 20K_GINB and 20K_RNDM gave the highest correlation (Fig. 4). The number of animals of a particular breed in the core had a large influence

on the consistency with non-APY for that breed. Only RNDM, OLD and YNG cores included Texel and Highlander animals, and 5K_YNG had only 1 Highlander animal in the core. This shows the importance of including minor breeds in the core sample; otherwise, the consistency with non-APY would decrease for those breeds.

Figure 5 presents bias from non-APY for Texel and Highlander genotyped animals. The 20K_RNDM and 20K_GINB analyses showed low bias. For the genotyped Highlander animals, all the 5K cores had low correlation and high bias, and all the 20K cores had high correlation and low bias (Figs 4 and 5).

Some G^{-1} Properties and Time for Solving the Equations

The difference between the non-APY and the APY evaluations was that a sparser (proportional to the number of noncore animals) G_{APY}^{-1} was used

Figure 4. Correlations between the non-APY GEBV and the APY GEBV using different cores, for Texel and Highlander genotyped animals^a. ^assGBLUP = single-step GBLUP, CP = Coopworth-Perendale, ROM = Romney, COMP = Composite, OLD = Oldest, YNG = Youngest, AINB = most inbred in $\mathbf{A}_{22}$ (the block of the pedigree relationship matrix corresponding to genotyped animals), GINB = most inbred in $\mathbf{G}$ (genomic relationship matrix), RNDM = Random.

rather than $\mathbf{G}^{-1}$. In APY, $\mathbf{G}^{\text{nn}}$ ($\mathbf{G}^{-1}$ block for noncore animals) is the diagonal matrix $\mathbf{M}_{nn}^{-1}$. According to [Ostersen et al., 2016](#), sparser $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$ reduced the computational load and converged faster. The sparsity of the off-diagonal elements of $\mathbf{G}_{\text{APY}}^{-1} - \mathbf{A}_{22}^{-1}$ was in a range of 56.5 to 57.6% for 5K cores, 42.7 to 45.2% for 10K core, and 22.1 to 24.5% for 20K cores. Within core size, breed-based cores had the lowest, and AINB had the maximum sparsity. In non-APY, 1,701 of 1,102,080,826 off-diagonal elements of $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$ were zero.

[Table 3](#) shows the time and the number of PCG iterations spent to solve the MME for the APY analyses with different cores. The number of PCG iterations and average time per iteration increased by increasing the core size. Generally, RNDM converged faster and breed-based cores converged slower than the other analyses. Also, OLD converged faster than YNG. Solving time and the number of PCG iterations were not only dependent on the sparsity of $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$. For example, non-APY converged faster than some of the APY analyses, or some APY analyses converged faster or in less number of iterations compared with 5KCP, even though they had a larger core size ([Table 3](#)). BLUP converged in 00:46:21 time and 668 PCG iterations, and non-APY converged in 14:41:40 time and 733 PCG iterations. The server specifications are given in [Table 3](#).

Eigenvalue decomposition of $\mathbf{G}$ can be used to assess the optimum number of core animals ([Misztal,](#)

[2016](#)). The highest 5K, 10K, 15K, and 20K eigenvalues explained 81.2, 91.9, 96.3, and 98.4% of the variation in $\mathbf{G}$, respectively. The influence of the core sample on the distribution of the diagonal elements of $\mathbf{G}^{-1}$ is given in [Fig. 6](#). The variation of the diagonal values was reduced by APY, more so with smaller cores. Within each core size, OLD produced the largest, and YNG produced the lowest variation of the $\mathbf{G}^{-1}$ diagonal values. Descriptive statistics of GEBV for genotyped animals are presented in [Supplementary Table 1](#). The EBV from BLUP was in a range of -4.062 and 9.534, with a mean and SD of 2.216 ± 1.425 . The GEBV showed a larger SD, with averages closer to 0, compared with EBV.

DISCUSSION

APY reduces the computational cost, and improves the prediction accuracy of GEBV ([Lourenco et al., 2015](#); [Fragomeni et al., 2015a, 2015b](#); [Misztal, 2016](#); [Masuda et al., 2016](#); [Ostersen et al., 2016](#); [Pocrnic et al., 2016a, 2016b](#); [Bradford et al., 2017](#)). It makes use of a sparser $\mathbf{G}^{-1}$ matrix ($\mathbf{G}_{\text{APY}}^{-1}$) compared with the regular $\mathbf{G}^{-1}$. The off-diagonals of a partition of $\mathbf{G}^{-1}$ that has no contribution to the accuracy, and contains redundant information are ignored ([Misztal, 2016](#)). Consequently, due to the relationships between noncore animals and nongenotyped ancestors, applying APY provides GEBV approximations to the non-APY GEBV. In

Figure 5. Regression coefficients (slopes) of the APY GEBV using different cores on the non-APY GEBV, for Texel and Highlander animals^a. ^assGBLUP = single-step GBLUP, CP = Coopworth-Perendale, ROM = Romney, COMP = Composite, OLD = Oldest, YNG = Youngest, AINB = most inbred in A_{22} (the block of the pedigree relationship matrix corresponding to genotyped animals), GINB = most inbred in G (genomic relationship matrix), RNDM = Random.

APY, genomic inference for noncore animals are provided by the coefficients from core animals, and GEBV of noncore animals is conditional on the GEBV of core animals.

Effective Dimensionality of G

The effective dimensionality of G is equal to the rank of G , or the number of independent chromosome segments in the population (ICS; Misztal, 2016). Number of core animals equal to the effective

dimensionality of G is sufficient to account for all the information, assuming that the segregation of all the ICS is present in the core animals. Using less core animals than the number of ICS would reduce the accuracy and using more core animals would not be beneficial and may even decrease the numerical stability in the MME (Misztal, 2016). An optimal number of core animals can be determined by assessing the accuracy of GEBV with APY and varied cores, or by eigenvalue decomposition of G (Misztal, 2016; Pocrnic et al., 2016a, 2016b;

Table 3. Number of preconditioned conjugate gradient iterations and time (hh:mm:ss) spent to reach convergence (convergence criterion = $1.0\text{e-}12$) for the APY analyses^a

Core ^b	5,000	10,000	20,000
CP	1,433 (13:52:07)		
ROM	1,244 (11:39:03)	1,644 (18:21:17)	
COMP	1,114 (09:18:08)	1,364 (13:42:35)	1,789 (22:26:37)
OLD	796 (07:53:13)	867 (11:13:13)	1,037 (15:57:11)
YNG	1,100 (10:34:02)	1,106 (12:36:32)	1,237 (17:11:01)
AINB	883 (08:13:50)	944 (09:43:33)	1,133 (15:29:08)
GINB	817 (06:31:33)	850 (08:24:50)	994 (15:44:38)
RNDM	730 (06:56:04)	764 (07:44:19)	803 (10:20:11)

^aSolving BLUP and non-APY equations took 00:46:21 and 14:41:40 time, in 668 and 733 iterations, respectively; Machine specifications: 4×Intel Xeon Processors X7550, 18MB Cache, 2.00 GHz, 1TB RAM, 32 physical (64 logical) cores.

^bCP = Coopworth-Perendale, ROM = Romney, COMP = Composite, OLD = Oldest, YNG = Youngest, AINB = most inbred in A_{22} (the block of the pedigree relationship matrix corresponding to genotyped animals), GINB = most inbred in G (genomic relationship matrix), RNDM = Random.

Figure 6. Distributions (boxplots) of $\text{diag}(\mathbf{G}^{-1})$ from the non-APY and the APY analyses using different cores^a. ^aCore size_type, with core types being CP = Coopworth-Perendale, ROM = Romney, COMP = Composite, OLD = Oldest, YNG = Youngest, AINB = most inbred in A_{22} (the block of the pedigree relationship matrix corresponding to genotyped animals), GINB = most inbred in G (genomic relationship matrix), RNDM = Random.

Bradford et al., 2017). Pocrnic et al. (2016a, 2016b) concluded that the effective dimension of G is a function of effective population size (N_e), and the largest eigenvalues explaining 90, 95, and 98% of the variation in G , correspond to $\sim N_e L$, $\sim 2N_e L$, and $\sim 4N_e L$, where L is the genome length in Morgans. They recommended using a core size that would explain 98% of the variation in G . The last 2% of the variation introduces noise, probably due to the segregation of noncoding regions. As a result, GEBV from some APY analyses can be a better estimator of true breeding value than GEBV from the non-APY analysis. This has been shown in simulated data by Pocrnic et al., 2016a, and validated in real data by Pocrnic et al., 2016b. In our study, the number of the highest eigenvalues explaining 90, 95, and

98% of the variation in G were 8,625, 13,123, and 18,798, respectively, showing the ratios of 1, 1.5, and 2.2. These ratios were not in agreement with the ratios of 1, 2, and 4 reported by Pocrnic et al. (2016a, 2016b). A recent study (Priour et al., 2017) has predicted N_e for different breeds and crosses of sheep in New Zealand, using different methods and considering different number of historical generations.

Core Size and Incoming Genotypes

APY is mainly of interest when the number of genotypes makes the direct inversion of G infeasible or impractical. Direct inversion of G was feasible with the number of genotypes in our study. However, like most livestock populations,

the number of genotypes is rapidly increasing in sheep. We tested the performance of APY before facing those limitations. Assuming adding new genotypes does not cause a sudden increase in N_e , the efficient number of core animals found in this study would be enough for more genotypes, as $\mathbf{G}$ has a limited dimensionality (Pocrnic et al., 2016a, 2016b). However, additional genotypes can increase N_e when the number of new genotypes is relatively large (Pocrnic et al., 2016b), or if the new genotyped animals are from a diverse genetic background (e.g., a different breed) or have been under a different selection program compared with the previously genotyped population. Another factor that may increase the number of core animals required is the number of markers used. Increasing genotype density and whole genome sequencing can reveal more ICS.

Core Size and Type

A core size of 10K was insufficient for the population under the study as in the best situation (10K_RNDM) the GEBV correlations with non-APY were 0.990 and 0.964 for core and noncore (0.969 for all the genotyped) animals, respectively. A core size of 20K performed well and 20K_RNDM got a correlation > 0.99 , for all the 3 groups of animals. However, core type was important, for example, 20K_COMP performed worse than 10K_RNDM (Figs 2 and 3). There are contrary reports on the importance of core type. Some studies (Lourenco et al., 2015; Ostensen et al., 2016) found it important, while other studies (Fragomeni et al., 2015a, 2015b; Bradford et al., 2017) reported similar GEBV from APY of different subset of core animals of the same core size. Bradford et al. (2017) mentioned that core type is important, when genotyped animals have unknown parents. In our studied population, only 4.7 and 0.2% of the genotyped animals had unknown sire and unknown dam, respectively, and no genotyped animals were missing both parents. Considering genotyped parents, 20.7, 94.9, and 20.0% of the genotyped animals had no genotyped sire, no genotyped dam, and neither, respectively. If a core contains animals from a specific breed, the influence on the other breeds would depend on how close or diverse their genetic makeup is. Genotyping more animals from minor breeds can reduce the deficiency for those breeds. No matter whether APY is applied or not, using high-density SNP chips can improve multibreed genomic evaluations.

The GEBV from some APY analyses might be more accurate than the GEBV from the non-APY analysis. Assuming that 2% of the variation in $\mathbf{G}$ is associated with noise (depending on the quality of $\mathbf{G}$) and reducing GEBV accuracies (Pocrnic et al., 2016a), an APY with 18.8K core animals could possibly result in more accurate GEBV than the GEBV from the non-APY analysis, depending on the core type.

There was not much gain to obtain beyond 20K core animals in terms of higher consistency with non-APY and lower bias from non-APY. The correlation between non-APY and 20K_RNDM GEBV was greater than 0.99. An APY analysis with a 15K_RNDM core was performed and it could also achieve correlations > 0.995 for nongeno and core animals, and > 0.98 for non-core animals. Other studies have investigated core sizes in other species. The required core size was 10K for American Angus (Lourenco et al., 2015) and American Holstein cattle (Fragomeni et al., 2015a; Masuda et al., 2016), and 4.5K to 6.5K for Danish pigs (Ostensen et al., 2016). Pocrnic et al. (2016b) reported optimal results obtained with 14K core animals for Holstein and Angus cattle, 12K for Jersey cattle, and 6K for pigs and broiler chickens. However, they reported minimal accuracy loss, when half of these numbers were used. Evaluation in a multibreed context might mean a greater required number of core animals. This is consistent with an expected increase in N_e , which also implies greater linkage disequilibrium and ICS. Accuracy of APY increased more from 5K to 10K than from 10K to 20K core animals. Other studies reported similar observations (Lourenco et al., 2015; Fragomeni et al., 2015a; Ostensen et al., 2016; Pocrnic et al., 2016a, 2016b). The reason is that most of the variation in $\mathbf{G}$ is explained by the largest eigenvalues. The addition of more core animals corresponds to smaller additional eigenvalues adding marginal accuracy. Also, adding more genotypes, less additional core animals would be needed (Ostensen et al., 2016).

Similar to other studies (Lourenco et al., 2015; Fragomeni et al., 2015a, 2015b; Masuda et al., 2016; Ostensen et al., 2016; Bradford et al., 2017), we found the RNDM cores to be the most beneficial, which is due to multiple generations being represented in the core (Ostensen et al., 2016; Bradford et al., 2017). For the same reason, OLD and YNG did not perform well, and it took them more time and number of PCG iterations for convergence (Table 3). In another study, the slowest convergence was found using the youngest animals in the core

(Bradford et al., 2017). Compared with the other APY analyses, the RNDM cores converged in the lowest number of iterations (Table 3), which according to Misztal (2016) is due to a better numerical conditioning. Using ssGBLUP with APY inverse of G , resulted in more stable and faster convergence in purebred pigs while not reducing the accuracies (Pocrnic et al., 2017). The non-APY analysis converged in 733 iterations, similarly Ostersen et al. (2016) reported fewer PCG iterations needed for the non-APY analysis than for most of their APY analyses. However, non-APY had the highest average time per iteration.

We found low consistency with non-APY and high bias from non-APY for Texel and Highlander animals, especially when they had no representative animals in the core (Figs 4 and 5). It is important to include animals from different breed compositions and all the genotyped generations in the core sample. The RNDM cores were superior, because they best fulfilled these conditions. Studying a simulated data of a 3-way crossbreeding program, Vandenplas et al. (2018) reported that APY GEBV approximations of non-APY GEBV were accurate when the core contained animals from different breeds and crosses, and the core size was equal to the number of the largest eigenvalues explaining between 98 and 99% of the variation in G . Ostersen et al. (2016) recommended cores representing all generations and maximizing the number of genotyped offspring, because parents represent ICS of their offspring. They obtained such a sample from a random sample from the last generation of genotyped animals and a sample from other generations of genotyped animals with the highest number of genotyped offspring. Small accuracy gains were achieved over RNDM for Danish pigs. In our population, only 11.7% of the genotyped animals had a genotyped offspring, and 8.1% had more than one genotyped offspring. Therefore, choosing animals based on the number of genotyped offspring could not provide enough number of core animals. There might be a trade-off between the randomness of the core sample representing the genotyped population and the number of genotyped offspring per genotyped animal. This potentially could favor specific breeds, generations, flocks, or families, which would result in a worse performance.

Concluding Remarks

APY can be successfully applied to ssGBLUP evaluations of livestock to deliver accurate results at lower computational expenses. For accurate

APY evaluations, the number of core animals and the core type are the key factors. Inadequate core size and improper core types can cause accuracy loss and bias. The consequences are not limited to noncore animals, but also core and nongenotyped animals. Random cores proved to perform well due to a better coverage of the generations and breeds. Finding breed-generation proportions in the genotyped population and nonrandom sampling from each into the core might be a possibility. However, given the performance of random samples, further gain would be very marginal.

SUPPLEMENTARY DATA

Supplementary data are available at *Journal of Animal Science* online.

Conflict of interest statement. None declared.

LITERATURE CITED

- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743–752. doi:10.3168/jds.2009–2730
- Bradford, H. L., I. Pocrnić, B. O. Fragomeni, D. A. L. Lourenco, and I. Misztal. 2017. Selection of core animals in the Algorithm for Proven and Young using a simulation model. *J. Anim. Breed. Genet.* 134:545–552. doi:10.1111/jbg.12276
- Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:2. doi:10.1186/1297-9686-42-2
- Fragomeni, B. O., D. A. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015a. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *J. Dairy Sci.* 98:4090–4094. doi:10.3168/jds.2014-9125
- Fragomeni, B. O., D. A. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, and I. Misztal. 2015b. Use of genomic recursions and algorithm for proven and young animals for single-step genomic BLUP analyses—a simulation study. *J. Anim. Breed. Genet.* 132:340–345. doi:10.1111/jbg.12161
- Henderson, C. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics.* 32:69–83. doi:10.2307/2529339
- Lourenco, D. A., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, et al. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American angus. *J. Anim. Sci.* 93:2653–2662. doi:10.2527/jas.2014-8836
- Masuda, Y., I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar, D. A. L. Lourenco, B. O. Fragomeni, and T. J. Lawlor. 2016. Implementation of genomic recursions in single-step genomic best linear unbiased

- predictor for US Holsteins with a large number of genotyped animals. *J. Dairy Sci.* 99:1968–1974. doi:10.3168/jds.2015-10540
- Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202:401–409. doi:10.1534/genetics.115.182089
- Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci.* 97:3943–3952. doi:10.3168/jds.2013-7752
- Ostensen, T., O. F. Christensen, P. Madsen, and M. Henryon. 2016. Sparse single-step method for genomic evaluation in pigs. *Genet. Sel. Evol.* 48:48. doi:10.1186/s12711-016-0227-8
- Pocrnic, I., D. A. L. Lourenco, H. L. Bradford, C. Y. Chen, and I. Misztal. 2017. Technical note: Impact of pedigree depth on convergence of single-step genomic BLUP in a purebred swine population. *J. Anim. Sci.* 95:3391–3395. doi:10.2527/jas.2017.1581
- Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. *Genetics* 203:573–581. doi:10.1534/genetics.116.187013
- Pocrnic, I., D. A. Lourenco, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. *Genet. Sel. Evol.* 48:82. doi:10.1186/s12711-016-0261-6
- Prieur, V., S. M. Clarke, L. F. Brito, J. C. McEwan, M. A. Lee, R. Brauning, K. G. Dodds, and B. Auvray. 2017. Estimation of linkage disequilibrium and effective population size in New Zealand sheep using three different methods to create genetic maps. *BMC Genet.* 18:68. doi:10.1186/s12863-017-0534-2
- Qiu, Y., and J. Mei. 2018. RSpectra: Solvers for large-scale eigenvalue and SVD problems. R package version 0.13–1–[accessed November 20, 2018]. <https://CRAN.R-project.org/package=RSpectra>
- Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* 32:949–953. doi:10.2307/2529279
- R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15:478. doi:10.1186/1471-2164-15-478
- Strandén, I., and M. Lidauer. 1999. Solving large mixed models using preconditioned conjugate gradient iteration. *J. Dairy Sci.* 82:2779–2787. doi:10.1111/jbg.12257
- Vandenplas, J., M. P. L. Calus, and J. Ten Napel. 2018. Sparse single-step genomic BLUP in crossbreeding schemes. *J. Anim. Sci.* 96:2060–2073. doi:10.1093/jas/sky136
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. doi:10.3168/jds.2007-0980